ИНФОРМАТИКА, ВЫЧИСЛИТЕЛЬНАЯ ТЕХНИКА И УПРАВЛЕНИЕ

УДК 004.912
DOI 10.22213/2413-1172-2018-3-166-171

К ВОПРОСУ О ПОСТРОЕНИИ ИНФОРМАЦИОННОЙ СИСТЕМЫ ОБРАБОТКИ ТЕКСТОВОЙ ИНФОРМАЦИИ НА ОСНОВЕ СМЫСЛОВЫХ ГРУПП

С. В. Моченов, кандидат технических наук, профессор, ИжГТУ имени М. Т. Калашникова, Ижевск, Россия
М. В. Втюрин, аспирант, ИжГТУ имени М. Т. Калашникова, Ижевск, Россия
Р. Р. Ахметгалеев, аспирант, ИжГТУ имени М. Т. Калашникова, Ижевск, Россия

Рассматривается информационный подход к анализу текста, основанный на разбиении предложения на составные части и выделении темы и ремы. Актуальность работы обосновывается необходимостью поиска и выделения полезной для пользователя информации, которую он мог бы использовать при выполнении необходимых научных исследований. Введено понятие структурно-семантических смысловых групп предложений и определены требования к ним. Смысловая группа предложения определяется на основе анализа связей между словами предложения и включает в себя некоторый набор рядом расположенных слов, задающих некоторый новый образ. Описан сценарий анализа текстовой информации на основе предлагаемого подхода. Приведено описание подходов при разбиении текста документа на смысловые группы. Приведены развернутые результаты работы программного комплекса при различной целевой установке на обработку текста.

Представленные результаты показывают возможности разработанного программного комплекса: выполнение структуризации отдельных предложений текста; формирование ключевых слов в виде смысловых групп для дальнейшего анализа; отбор смысловых групп, определяющих основной смысл предложений и текста; значительное сокращение текста при сохранении смысловой составляющей. В дальнейшем предполагается расширение функциональных возможностей комплекса и проверка основных идей при обработке больших информационных массивов.

Ключевые слова: информационная система, обработка текстовой информации, смысловые группы, сокращение текста, смысловая составляющая, выделение информации.

Введение

Задачи, связанные с автоматизацией обработки текстовой информации, особенно актуальны в настоящее время, поскольку в мире наблюдается быстрый рост информационных потоков по различным направлениям человеческой деятельности.

Исследователь, занимающийся решением какой-либо проблемы, вынужден изучать уже накопленный опыт, который отражен в многочисленных текстах, хранящихся в библиотеках и в глобальной сети Интернет, для формирования своих научных результатов в той или иной области знаний. Обработка текстов заключается в поиске и выделении нужной и полезной для пользователя информации [1–4], которую он мог бы использовать при реализации собственных идей и выполнения необходимых научных исследований.

Автор какой-либо научной статьи (публикации) раскрывает определенные аспекты своих исследований, опираясь на собственный опыт и тот научный базис, который используется им в процессе научной деятельности. Реализация авторского замысла осуществляется в виде структурно и логически организованного текста научной публикации. Исследователь (читатель, пользователь) при анализе той или иной статьи ориентируется на свои интересы, которые связаны с его научной или практической деятельностью.

Учитывая то, что и автор статьи, и пользователь-исследователь работают в своих информационных полях конкретных знаний, возникает вопрос, как автоматически должна быть обработана текстовая информация, чтобы уменьшить ее объем и одновременно сохранить основную идею, смысл основных положений, изложенных

автором статьи, для предъявления этой информации пользователю.

Любая статья, научная публикация представляет некоторые результаты, которые хотел бы отобразить автор. Можно говорить о некоторой функции цели написания статьи: конечная цель текста как смысловая совокупность результатов складывается из промежуточных целей, которые обусловлены структурой отдельных предложений, абзацев, разделов и т. д. Налицо иерархическая организация структуры текста.

Описание методов исследования

В работе [5] было введено понятие вектора цели. Координатное пространство вектора цели на каждом уровне иерархии текста имеет свою метрику. На уровне отдельных предложений эта метрика определяется смысловым содержанием предложения. Вектор цели предложения определяется пространством структурно-семантических смысловых групп (СГ).

Смысловая группа предложения определяется на основе анализа связей между словами предложения и включает в себя некоторый набор рядом расположенных слов, задающих некоторый новый образ. Смысловая характеристика этого образа, его значение, определяется смыслом входящих в этот образ слов. Таким образом, любое предложение можно рассматривать как некоторое связное множество образов, определяемых через смысловые группы. При этом элементы, составляющие СГ, связаны через соответствующую грамматическую основу.

Количество смысловых групп определяется сложностью анализируемого предложения, а само предложение представляется как новый образ, обобщенное слово в пространстве образов смысловых групп, выражающих некоторую законченную мысль. Последовательность предложений текста задает соответствующую иерархию образов и соответственно координат в пространстве целей.

Структуризация отдельных предложений как самостоятельных элементов смысла предполагает наличие процедуры разбиения предложения на некоторые самостоятельные смысловые единицы, смысловые группы.

С учетом вышесказанного основным требованием к СГ является требование отображения смысловой законченности, способности выражать некоторый смысл более высокого порядка по сравнению со смыслом отдельных слов, составляющих СГ.

В ряде работ [6–10] рассматривается перспективный, по нашему мнению, информационный подход к анализу текста, основанный на разбиении предложения (а в дальнейшем и всего текста) на составные части – тему и рему. Обычно дается интерпретация темы предложения как некоторого данного (исходного), а ремы – как чего-то нового. Тема отвечает на вопрос, о чем говорится, а рема – на вопрос, что говорится. При анализе отдельных предложений текста разбиение на тему и рему можно рассматривать как первый этап структуризации предложения.

В данной работе в качестве разделителя на тему и рему в предложении выступает первый встреченный глагол. При этом левая часть от глагола считается темой, а правая – ремой. Как левая, так и правая части могут содержать определенное количество слов. В левой части глаголы отсутствуют, а в правой части наряду с другими элементами предложения может оказаться несколько глаголов, что определяется сложностью предложения. Возможны различные варианты формирования предложения автором текста, в том числе и такие, когда глагол располагается в конце предложения. В этом случае левая часть, определяющая тему, будет представлять собой пустое множество слов. В дальнейшем разбиение предложения на смысловые группы осуществляется отдельно по теме и по реме.

При разбиении на смысловые группы используется ряд подходов, основанных на выделении отдельных элементов и сочетания элементов предложения. В роли таких элементов выступают предлоги, союзы, знаки препинания, вспомогательные слова, взаимное расположение отдельных частей речи с предлогами и союзами. Для удобства реализации заложенных в информационную систему алгоритмов выполняется индексация всех элементов предложения. Реализуется модульный принцип построения отдельных функций, выполняемых программным комплексом.

Сценарий анализа представляет собой следующий набор процедур.

1. Предварительная обработка текста.
2. Разбиение предложений по частям речи.
3. Разбиение предложений на тему и рему.
4. Формирование массивов элементов и индексов, необходимых для выделения смысловых групп.
5. Формирование смысловых групп.
6. Формирование результатов анализа.

Полученные данные в ходе экспериментального исследования и их интерпретация

В примерах 1–4 приведены развернутые результаты работы программного комплекса при различной целевой установке на обработку текста.

Программный комплекс как информационная система реализован с использованием среды разработки JetBrains PyCharm Community Edition 2017.3 x64, Python36-32.

Пример 1. Разделение сложного предложения на смысловые группы простых предложений.

C:\Users\стас\AppData\Local\Programs\Python\Python36-32\python.exeC:/Users/стас/Desktop/ру3eg/1_12_2.py

Просим вас разобраться в сложившейся ситуации и принять меры к недобросовестным исполнителям работ по капитальному ремонту нашего дома и потребовать от них обязательного утепления чердака по всему периметру.

[] – тема.

['просим', 'вас', 'разобраться', 'в', 'сложившейся', 'ситуации', 'и', 'принять', 'меры', 'к', 'недобросовестным', 'исполнителям', 'работ', 'по', 'капитальному', 'ремонту', 'нашего', 'дома', 'и', 'потребовать', 'от', 'них', 'обязательного', 'утепления', 'чердака', 'по', 'всему', 'периметру', '.'] – рема.

[0, 2, 7, 19] – массив индексов глаголов в реме, упорядоченный массив.

['и'].

['и', 'и'].

['и', 'и', '.'].

['и', 'и', '.'] – массив предлогов, союзов и вспомогательных слов в реме. Формируется в процессе анализа.

[6, 18, 28] – массив индексов предлогов, союзов и вспомогательных слов в реме.

3 – число предлогов, союзов и вспомогательных слов в реме.

Смысловые группы простых предложений:

['просим', 'вас', 'разобраться', 'в', 'сложившейся', 'ситуации']

['и', 'принять', 'меры', 'к', 'недобросовестным', 'исполнителям', 'работ', 'по', 'капитальному', 'ремонту', 'нашего', 'дома']

['и', 'потребовать', 'от', 'них', 'обязательного', 'утепления', 'чердака', 'по', 'всему', 'периметру'].

Process finished with exit code 0

Пример 2. Смысловые группы ремы, полученные на основе анализа массивов индексов.

C:\Users\стас\AppData\Local\Programs\Python\Python36-32\python.exeC:/Users/стас/Desktop/ру3eg/1_12_2.py

Просим вас разобраться в сложившейся ситуации и принять меры к недобросовестным исполнителям работ по капитальному ремонту нашего дома и потребовать от них обязательного утепления чердака по всему периметру.

[] – тема.

['просим', 'вас', 'разобраться', 'в', 'сложившейся', 'ситуации', 'и', 'принять', 'меры', 'к', 'недобросовестным', 'исполнителям', 'работ', 'по', 'капитальному', 'ремонту', 'нашего', 'дома', 'и', 'потребовать', 'от', 'них', 'обязательного', 'утепления', 'чердака', 'по', 'всему', 'периметру', '.'] – рема.

[0, 2, 7, 19] – массив индексов глаголов в реме, упорядоченный массив.

['в']

['в', 'и']

['в', 'и', 'к']

['в', 'и', 'к', 'по']

['в', 'и', 'к', 'по', 'и']

['в', 'и', 'к', 'по', 'и', 'от']

['в', 'и', 'к', 'по', 'и', 'от', 'по']

['в', 'и', 'к', 'по', 'и', 'от', 'по', '.']

['в', 'и', 'к', 'по', 'и', 'от', 'по', '.'] – массив предлогов, союзов и вспомогательных слов в реме. Формируется в процессе анализа.

[3, 6, 9, 13, 18, 20, 25, 28] –массив индексов предлогов, союзов и вспомогательных слов в реме.

8 – число предлогов, союзов и вспомогательных слов в реме.

Смысловые группы ремы:

['просим', 'вас', 'разобраться']

['в', 'сложившейся', 'ситуации']

['и', 'принять', 'меры']

['к', 'недобросовестным', 'исполнителям', 'работ']

['по', 'капитальному', 'ремонту', 'нашего', 'дома']

['и', 'потребовать']

['от', 'них', 'обязательного', 'утепления', 'чердака']

['по', 'всему', 'периметру']

Process finished with exit code 0

Пример 3. Смысловые группы ремы, полученные на основе анализа индексов предлогов, союзов и вспомогательных слов

C:\Users\стас\AppData\Local\Programs\Python\Python36-32\python.exeC:/Users/стас/Desktop/py3eg/1_12_2.py

Просим вас разобраться в сложившейся ситуации и принять меры к недобросовестным исполнителям работ по капитальному ремонту нашего дома и потребовать от них обязательного утепления чердака по всему периметру.

[] – тема.

['просим', 'вас', 'разобраться', 'в', 'сложившейся', 'ситуации', 'и', 'принять', 'меры', 'к', 'недобросовестным', 'исполнителям', 'работ', 'по', 'капитальному', 'ремонту', 'нашего', 'дома', 'и', 'потребовать', 'от', 'них', 'обязательного', 'утепления', 'чердака', 'по', 'всему', 'периметру', '.'] – рема.

[0, 2, 7, 19] – массив индексов глаголов в реме, упорядоченный массив.

['разобраться']
['разобраться', 'и']
['разобраться', 'и', 'принять']
['разобраться', 'и', 'принять', 'к']
['разобраться', 'и', 'принять', 'к', 'по']
['разобраться', 'и', 'принять', 'к', 'по', 'и']
['разобраться', 'и', 'принять', 'к', 'по', 'и', 'потребовать']
['разобраться', 'и', 'принять', 'к', 'по', 'и', 'потребовать', 'обязательного']
['разобраться', 'и', 'принять', 'к', 'по', 'и', 'потребовать', 'обязательного', 'по']
['разобраться', 'и', 'принять', 'к', 'по', 'и', 'потребовать', 'обязательного', 'по', '.']
['разобраться', 'и', 'принять', 'к', 'по', 'и', 'потребовать', 'обязательного', 'по', '.'] – массив предлогов, союзов и вспомогательных слов в реме.

[2, 6, 7, 9, 13, 18, 19, 22, 25, 28] – массив индексов предлогов, союзов и вспомогательных слов в реме.

10 – число предлогов, союзов и вспомогательных слов в реме.

Смысловые группы ремы:

['просим', 'вас'] ['разобраться', 'в', 'сложившейся', 'ситуации'] ['и'] ['принять', 'меры'] ['к', 'недобросовестным', 'исполнителям', 'работ'] ['по', 'капитальному', 'ремонту', 'нашего', 'дома'] ['и'] ['потребовать', 'от', 'них'] ['обязательного', 'утепления', 'чердака'] ['по', 'всему', 'периметру'].

Process finished with exit code 0

Пример 4. Сокращение объема предложения на основе выделения значимых смысловых групп.

C:\Users\стас\AppData\Local\Programs\Python\Python36-32\python.exeC:/Users/стас/Desktop/py3eg/1_12_6.py

Просим вас разобраться в сложившейся ситуации и принять меры к недобросовестным исполнителям работ по капитальному ремонту нашего дома и потребовать от них обязательного утепления чердака по всему периметру.

['в', 'к', 'по', 'от', 'по'] – массив предлогов всего предложения – Words1_0 predlogi.

[0, 7, 2, 19] – массив индексов глаголов по всему предложению.

4 – число индексов глаголов.

[]–тема.

['просим', 'вас', 'разобраться', 'в', 'сложившейся', 'ситуации', 'и', 'принять', 'меры', 'к', 'недобросовестным', 'исполнителям', 'работ', 'по', 'капитальному', 'ремонту', 'нашего', 'дома', 'и', 'потребовать', 'от', 'них', 'обязательного', 'утепления', 'чердака', 'по', 'всему', 'периметру', '.'] – рема.

[0, 2, 7, 19] – массив индексов глаголов в реме, упорядоченный массив.

['в', 'и', 'к', 'по', 'и', 'от', 'по', '.'] – массив предлогов, союзов и вспомогательных слов в реме. Формируется в процессе анализа.

[1, 1, 1, 2, 1, 2, 2, 1] – список длин предлогов, союзов и вспомогательных слов в реме.

[3, 6, 9, 13, 18, 20, 25, 28] – массив индексов предлогов, союзов и вспомогательных слов в реме.

8 – число предлогов, союзов и вспомогательных слов в реме.

Смысловые группы ремы. Выделение смысловых групп на основе анализа массивов индексов предлогов, союзов и вспомогательных слов ремы в соответствии с требованиями экспертов:

['просим', 'вас']
['разобраться', 'в', 'сложившейся', 'ситуации']
['и']
['принять', 'меры']
['к', 'недобросовестным', 'исполнителям', 'работ']
['по', 'капитальному', 'ремонту', 'нашего', 'дома']
['и']
['потребовать', 'от', 'них']
['обязательного', 'утепления', 'чердака']
['по', 'всему', 'периметру']

Значимые смысловые группы ремы:

(['просим', 'вас', 'разобраться'], ['и'], ['принять', 'меры']).

Process finished with exit code 0

Выводы

Таким образом, представленные результаты показывают возможности разработанного программного комплекса: выполнение структуризации отдельных предложений текста; формирование ключевых слов в виде смысловых групп для дальнейшего анализа; отбор смысловых групп, определяющих основной смысл предложений и текста; значительное сокращение текста при сохранении смысловой составляющей. В дальнейшем предполагается расширение функциональных возможностей комплекса и проверка основных идей при обработке больших информационных массивов.

Библиографические ссылки

1. *Алексеев А. А.* Тематический анализ новостного кластера как основа для автоматического аннотирования // Программная инженерия. 2014. № 3. С. 41–48.

2. *Артюхин В. В.*, *Чяснавичюс Ю. К.* Планирование аналитического исследования при помощи методов анализа качественных данных // Прикладная информатика. 2014. № 2. С. 23–48.

3. *Герте Н. А.*, *Курушин Д. С.*, *Нестерова Н. М.* Моделирование понимания текста как основа автоматизированного реферирования // Материалы VII Междунар. науч. конф. «Индустрия перевода» (Россия, Пермь, 1–3 июня 2015 г.). С. 81–84.

4. *Бледнов А. М.*, *Моченов С. В.*, *Луговских Ю. А.* Об одном методе статистической фильтрации текстовой информации // Современные информационные технологии и письменное наследие: от древних рукописей к электронным текстам : материалы междунар. науч. конф. (Россия, Ижевск, 13–17 июля 2006 г.). С. 126–130.

5. *Бледнов А. М.*, *Моченов С. В.*, *Луговских Ю. А.* Векторная модель представления текстовой информации // Современные информационные технологии и письменное наследие от древних рукописей к электронным текстам : материалы междунар. науч. конф. (Ижевск, 13–17 июля 2006 г). С. 136–145.

6. *Rankel P.*, *Conroy J.*, *Dang H.*, *Nenkova A.* A Decade of Automatic Content Evaluation of News Summaries: Reassessing the State of the Art // Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics. 2013. Pp. 131–136.

7. *Захарова И. С.*, *Филиппова Л. Я.* Основы информационно-аналитической деятельности : учебное пособие. Киев : Центр учебной литературы, 2013. 336 с.

8. *Курушин Д. С.*, *Нестерова Н. М.*, *Овчинникова И. Г.* О возможном подходе к созданию системы автоматического реферирования // Вопросы психолингвистики. 2014. № 2(20).С. 123–128.

9. [Abstracts - The Writing Center]. URL: http://writingcenter.unc.edu/handouts/abstracts/ (дата обращения: 02.04.2018).

10. *Осипов Г. С.*, *Шелманов А. О.* Метод повышения качества синтаксического анализа на основе взаимодействия синтаксических и семантических правил // Труды Шестой Междунар. конф. «Системный анализ и информационные технологии». 2015. Т. 1. С. 229–240.

References

1. Alekseev A. A. [Thematic analysis of a news cluster as a basis for automatic annotation]. *Programmnaya inzheneriya*, 2014, no. 3, pp. 41-48 (in Russ.).

2. Artjuhin V. V., Chjasnavichjus Ju. K. [Planning an analytical study using qualitative data analysis methods]. *Prikladnaja informatika*, 2014, no. 2, pp. 23-48 (in Russ.).

3. Gerte N. A., Kurushin D. S., Nesterova N. M. [Modeling the understanding of text as the basis for automated abstracting]. *Materialy VII Mezhdunar. nauch. konf.* "*Industriya perevoda*" (*Rossiya, Perm', 1-3 iyunya 2015 g.*) [Proc. VII of the Intern. Sci. Conf. "Translation Industry" (Russia, Perm, 1-3 June 2015), pp. 81-84 (in Russ.).

4. Blednov A. M., Mochenov S. V., Lugovskikh Yu. A. [About one method of statistical filtering of textual information]. *Materialy mezhdunar. nauch. konf. «Sovremennye informacionnye tehnologii i pis'mennoe nasledie: ot drevnih rukopisej k jelektronnym tekstam» (Rossija, Izhevsk, 13-17 ijulja 2006 g.)* [Proc. Intern. Sci. Conf. "Modern information technologies and written heritage: from ancient manuscripts to electronic texts" (Russia, Izhevsk, July 13-17, 2006), pp. 126-130 (in Russ.).

5. Blednov A. M., Mochenov S. V., Lugovskikh Yu. A. *Vektornaja model' predstavlenija tekstovoj informacii* [Vector model of text information representation]. *Materialy mezhdunar. nauch. konf. «Sovremennye informacionnye tehnologii i pis'mennoe nasledie ot drevnih rukopisej k jelektronnym tekstam» (Izhevsk, 13-17 ijulja 2006 g.)* [Proc. Intern. Sci. Conf. "Modern information technologies and written heritage from ancient manuscripts to electronic texts" (Izhevsk, July 13-17, 2006)], pp. 136-145 (in Russ.).

6. Rankel P., Conroy J., Dang H., Nenkova A. A Decade of Automatic Content Evaluation of News Summaries: Reassessing the State of the Art [Proc. 51-st Annual Meeting of the Association for Computational Linguistics]. 2013, pp. 131–136.

7. Zakharova I. S., Filippova L. Ya. *Osnovy informatsionno-analiticheskoi deyatel'nosti* [Fundamentals of Information and Analytical Activities]. Kiev, Tsentr uchebnoi literatury Publ., 2013, 336 p. (in Russ.).

8. Kurushin D. S., Nesterova N. M., Ovchinnikova I. G. [On a possible approach to the creation of an automatic referencing system]. *Voprosy psikholingvistiki*, 2014, no. 2(20), pp. 123-128 (in Russ.).

9. [Abstracts - The Writing Center]. Available at: http://writingcenter.unc.edu/handouts/abstracts/ (accessed 04.04.2018).

10. Osipov G. S., Shelmanov A. O. *Metod povyshenija kachestva sintaksicheskogo analiza na osnove vzaimodejstvija sintaksicheskih i semanticheskih*

pravil [Method for improving the quality of syntactic analysis based on the interaction of syntactic and semantic rules]. *Trudy Shestoj Mezhdunar. konf. «Sistemnyj analiz i informacionnye tehnologii»* [Proc. Sixth Intern. Conf. "System Analysis and Information Technologies"], 2015, vol. 1, pp. 229-240 (in Russ.).

To the Question of Developing an Information System for Processing Textual Information on the Basis of Semantic Groups

S. V. Mochenov, PhD in Engineering, Professor, Kalashnikov ISTU, Izhevsk, Russia
M. V. Vtyurin, Post-graduate, Kalashnikov ISTU, Izhevsk, Russia
R. R. Ahmetgaleev, Post-graduate, Kalashnikov ISTU, Izhevsk, Russia

The paper considers the information approach to the analysis of the text, based on splitting the sentence into its component parts and highlighting the topic and the rheme. The relevance of the work is justified by the need to search and identify useful information for the user, which he could use when performing the necessary scientific research. The concept of structural-semantic groups of sentences is introduced and requirements for them are defined. The sense group of a sentence is determined on the basis of an analysis of the relationships between the words of the sentence; and it includes some set of adjacent words that define some new image. The script of the analysis of the text information on the basis of the offered approach is described. A description of the approaches is provided for splitting the text of a document into semantic groups. The detailed results of the work of the program complex for various target installations for processing text are given.

The presented results show the capabilities of the developed software package: the structuring of individual text sentences; formation of key words in the form of semantic groups for further analysis; selection of semantic groups that determine the main meaning of sentences and text; significant reduction of the text while preserving the semantic component. In the future it is supposed to expand the functionality of the complex and check the main ideas when processing large information arrays.

Keywords: information system; processing of textual information; semantic groups; reduction of the text; semantic component, allocation of information.

Получено 20.04.2018